Helicobacter pylori's Seven-housekeeping Gene and *CagA* EPIYA Motif Patterns Linking to Gastrointestinal Diseases

Nattamon Narkwichearn[1], *Phataraporn* Khumphai[1], and *Sasiporn* Tongman[1*]

[1]Thammasat University, Faculty of Science and Technology, Klong Luang, Pathumthani, Thailand

Abstract. *Helicobacter pylori* (*H. pylori*) bacteria residing in human stomachs can cause gastrointestinal diseases and cancer. Discovering their effective sequences' biomarkers will help to estimate the disease risks. The *CagA* protein existing in some strains is one virulence factor. In this work, 272 *H. pylori* strain sequences were pulled from NCBI. Some types and patterns of *CagA* EPIYA motifs, including amino acid variations were only found in our study comparison with previous clinical data from literature. Two phylogenetic trees were built showing similar two main clades, one using *CagA* proteins translated by *cagA* genes and another using their concatenated seven-housekeeping genes. Our studied *CagA* protein set of EPIYA-ABD strains still report the same distribution of two deletion sites before the first EPIYA motif region in significance test. This aligns with the previous research, where their two-deletion was significantly discovered in EPIYA-ABD sequences isolated from gastric cancer patients. Moreover, the best alignment results, between seven allele sequences in each sequence type from pubMLST and seven housekeeping genes of the EPIYA-ABD strains, enable us to identify either EPIYA-ABD strain or strain groups. To conclude, several sequence analyses as in this work may further improve protocols in assessing the *H. pylori* gastric cancer risk.

1 Introduction

Helicobacter pylori (*H. pylori*), a class 1 carcinogen [1], is one of the factors in the development of gastrointestinal diseases, i.e. gastric ulcer, chronic stomach disease, duodenal cancer, primary B cell gastric lymphoma, and so on [2]. It can be transmitted between humans, and it is estimated that approximately 80% of the worldwide human's population infected by *H. pylori*. Besides, approximately 1-3% of infected patients are likely to develop into cancer [3]. Since, more than a half of all global patients infected by *H. pylori*, their genomes are so variant comparing with other bacteria [4]. Many methods are employed for classifying bacterial species, one is Multilocus Sequence Typing (MLST), which analyzes differences in each allele of the selected housekeeping genes and categorized them as sequence types (STs) [5]. The housekeeping genes are crucial for maintaining cell functions, because they must be consistently expressed for their entire life. The vital seven housekeeping

* Corresponding author: tongman.sas@gmail.com

genes of *H. pylori* used in the study of genetic diversity, evolution, and the inheritance of genetic characteristics includes *atpA, efp, ppa, mutY, trpC, ureI*, and *yphC* [6, 7]. They are responsible for different cellular functions as follows. First, *atpA* encodes ATP synthase sub-unit alpha, which synthesizes ATP for energy production. Second, *efp* produces a protein named elongation factor P, which prevents ribosome stalling during protein synthesis. Third, *ppa* encodes in-organic diphosphatase that participates in phosphate metabolism. Fourth, *mutY* codes for adenine DNA glycosylase which repairs DNA damage. Fifth, *trpC* yields indole-3-glycerol-phosphate synthase, a key enzyme in tryptophan biosynthesis. Sixth, *ureI* encodes a protein involved in acid-activated urea channel facilitating urea transport for stomach acid neutralization. Finally, *yphC* codes for GTPase, which aids in ribosome biogenesis and protein synthesis. Besides, two important genes, namely, *vacA* and *cagA* genes, relate to disease causation as virulence factors. The *vacA* gene is present in the genomes of all *H. pylori* strains [8]. But, *cagA* gene (cytotoxin-associated gene A) is situated in a *cag* pathogenicity island region (*cag*PAI) and is found in approximately 60% of all known *H. pylori* strains. The *cag*PAI comprises about 30 genes that encode type IV secretion system (T4SS) proteins, facilitating the transfer of *CagA* and peptidoglycan proteins [9]. *CagA* protein takes a part in tyrosine phosphorylation which is a process results in severe inflammation. This tyrosine phosphorylation happens in host cells at the *CagA* protein regions called the EPIYA motifs near the C-terminal end, and it is involved in pathogenic process. Roughly, EPIYA motifs can be classified into four types: A, B, C, and D [10]. Discovering important positions with variant form of *CagA* amino acid sequences near EPIYA motifs' regions can bring insightful about virulence level and gastrointestinal disease identification. For instance, *H. pylori* strains with EPIYA-ABD pattern, isolated from patients, are significantly found two-deletion positions locating before the *CagA* EPIYA motif type A region, especially, in cancer-causing variants [11]. Recently, *Helicobacter pylori* Genome Project (HpGP) as a resource of around one thousand high-quality genome sequences from clinical strains collected from worldwide was constructed. Their analysis discovered that *H. pylori* strains can be divided into subpopulations according to the human hosts' geographic origins [12]. So, many aspects of *H. pylori* sequence analyses associated with information about patient's gastric disease virulence will lead to accurate treatment and prevention.

Table 1. Four EPIYA motif types found on *CagA* amino acid sequences of *H. pylori*.

EPIYA Type	Amino Acid Pattern	Note	Reference
A	EPIYAKVNKKK	(A/T/V/S)GQ	Khaledi et al. [13]
B	EPIYAQVAKK	EPIY(A/T)(Q/K)	Khaledi et al. [13]
C	EPIYATIDDLG	-	Papadakos et al. [14]
D	EPIYATIDFDE	-	Papadakos et al. [14]

The variations of *CagA* amino acid sequence may contribute to disease severity. Understanding these variations can be an advantage in further studying for effective treatment. Therefore, combining clinical data with sequence analysis may enhance the knowledge about virulence and sequence patterns underlying disease pathogenesis which will help us in identifying key biomarkers useful for early diagnosis. In this research, initially, 272 *H. pylori* strains with complete genome sequence data were collected from NCBI database for our analysis. First, their *CagA* protein sequences were prepared. Next, EPIYA motif pattern of each strain was uncovered and compared with clinical evidences from many prior researches. Later, *CagA* protein sequences and seven-housekeeping gene were prepared for strain clustering and geographic distribution relationships via their two phylogenetic trees. Additionally, in our group of strains with EPIYA-ABD patterns comparing to this kind of group from the previous clinical research which is linked to patient's disease, the amino

acid polymorphism was analyzed. The variant form of two amino acid positions placing before the first EPIYA motif type A region of *CagA* protein was also observed. Apart from that, each sequence type (ST), containing seven allele sequences corresponding to each seven-housekeeping gene, from pubMLST database was aligned to our studied strains with EPIYA-ABD motif pattern. The exploration of using partial sequences inferring to the gastrointestinal disease risk was discussed.

2 Methods

2.1 *Helicobacter pylori* Clinical Research Collection and Genome Data

There are many clinical researches that studied *H. pylori* isolated from patients with various gastrointestinal disorders in term of molecular sequence analyses. Especially, virulence genes or protein sequence properties that might be strongly related to these diseases. Keikha and Karbalaei [15] collected literature to study the EPIYA motif genotypes (or patterns) and their correlation with severe clinical outcomes. In our work, four additional research articles about the patterns of *CagA* EPIYA motifs and their association with different clinical outcomes, i.e. general gastrointestinal diseases versus gastric cancer, were gathered from databases, such as ScienceDirect (www.sciencedirect.com), Google Scholar (https://scholar.google.com), and PubMed Central (www. ncbi.nlm.nih.gov/pmc).

To analyze *H. pylori* strains and their *CagA* EPIYA motif type and pattern, strains' sequences were prepared as follows. Based on *H. pylori* MLST scheme, a set of 2,740 sequence types (STs) was collected from pubMLST database (www.pubmlst.org) on July 6, 2022. The seven allele sequences of each ST were downloaded to blast against 744 *H. pylori* genome sequences from NCBI (www. ncbi.nlm.nih.gov) via standalone BLAST. The best hits with the highest percent identity of each ST were selected as 382 *H. pylori* strains in total. However, 60 strains were excluded because of their genome incompleteness. Then, the remaining 322 strains were divided into *cagA*-positive and *cagA*-negative groups referring to strains with and without *cagA*, respectively. Finally, only a group of 272 *cagA*-positive strains was our strain set used to further analyses. Denoting that there are different definitions about strain names among databases and research papers. In our research, text data storing in the "/strain=" feature qualifier associated with the "source" feature in "FEATURES" section of each genome GenBank file was pulled and became a strain name of each strain.

2.2 Seven Housekeeping Genes and *CagA* Protein Sequence Data Preparation

The complete gene sequences of seven housekeeping genes (*atpA, efp, mutY, ppa, trpC, ureI*, and *yphC*) and *CagA* protein sequences were extracted from 272 *H. pylori* complete genomes in the form of GenBank file format. Firstly, the starting and ending positions for each of seven housekeeping genes were located by using MAUVE program version 20150226 [16] in order to extract gene sequences. Then, complete sequences of seven genes for each strain were exported as separated FASTA files by writing in-house python scripts. Finally, locating the starting and ending translated protein positions of *cagA* gene in each GenBank file using the written python scripts and its basic function to filter *CagA* amino acid sequence and also extracting the information storing in "/country=" which is a part of the feature qualifiers in each GenBank genome file. The country information of the derived data file shows us about the source of each strain isolated from its human host living in different geographic area. But, only the small number of files contains no "/country=" feature qualifier.

2.3 *CagA* EPIYA Motif Pattern Identification

The python codes were written for detecting an EPIYA motif pattern on each of 272 *CagA* amino acid sequences. According to previous researches as Table 1, python functions relating to matching and splitting string tasks can be adopted to identify four EPIYA motif types, namely, A, B, C, and D. Denote that EPIYA motif types A and B contain some variations. For instance, the EPIYA motif type A is EPIYAKVNKKK in general case, however, in some cases, K after EPIYA can be replaced by A, T, V, or S, e.g. EPIYAAVNKKK and EPIYATVNKKK, or N can be replaced by G, e.g. EPIYAKVGKKK. Therefore, sequence matching process to identify all consecutive appearances of EPIYA types on *CagA* amino acid sequence results in EPIYA motif pattern. For example, EPIYA motif types A, B, and C were identified in order, hence, this is an EPIYA-ABC pattern.

2.4 *CagA* Amino Acid Polymorphism Study

In this study, *H. pylori* strain 26695 was used as a reference strain (NCBI accession: CP003904) as in Xue et al. (2021) amino acid polymorphism study. Every pairwise alignment of its *CagA* translated protein and each of 272 *H. pylori*'s *CagA* translated proteins was performed by writing a script using a "localalign" function from bioinformatics toolbox in MATLAB program version R2022a [17]. All two positions before the first EPIYA type appearing on *CagA* sequence were collected. The distribution of two-deletion positions, one-deletion position, and substitution position i.e. glutamic acid (*Glu*) to threonine (*Thr*) or glutamic acid (*Glu*) to asparagine (*Asn*), were analyzed.

2.5 Statistical Analysis

A chi-square test was conducted in order to investigate whether two frequencies come from the same distribution or not. These two frequencies are as follows. One was calculated by the observed occurrence numbers of two-deletion (and/or substitution) at two aligned positions of *CagA* proteins in the previous research dataset from Xue et al. (2021), and another was those occurrence numbers computed from our dataset. Due to limitation of clinical data source from Xue et al. (2021), their dataset was strain sequences with the EPIYA-ABD pattern isolated from patients. Our strain sequences pulled from NCBI database contains various patterns of the EPIYA motif, but only strains with EPIYA-ABD pattern were included in our dataset for comparison. R script was written to compute and analyze the *p*-value of chi-square test at 0.05 significant level using a command, chisq.test, from "stats" library.

2.6 Phylogenetic Tree Construction

A gene-by-gene multiple sequence alignment was performed to compare seven housekeeping genes among 272 *H. pylori* strains. For each strain, the aligned sequences of seven genes were concatenated into one FASTA file and used for phylogenetic tree construction. In addition, *CagA* amino acid sequences of 272 *H. pylori* strains were compared. Two phylogenetic trees based on concatenated sequences of seven housekeeping genes and based on *CagA* amino acid sequences were constructed by MEGA7 [18] and displayed by writing R script using "ggtree" library version 3.6.2 [19]. In MEGA7 set-up, each resulting multiple sequence alignment calculated by using ClustalW method [20]. The Maximum Composite Likelihood method [21] and the Neighbor-Joining algorithm [22] with the bootstrap method were used to estimate the reliable tree. Additionally, a bootstrap parameter was set to 1,000.

Table 2. Distribution of ten EPIYA motif patterns on *CagA* proteins with the occurrences of two main diseases (gastric disease and cancer) in *H. pylori* infected patients from collected clinical researches (see Section 2.1).

EPIYA Pattern	Reference	Number of Patients			Overall
		Gastric Disease	Cancer	Total	
AD	Xue et al. [11][8]	2	1	3	
	Chen et al. [26][4]	1	0	1	4
ABC	Beltrán-Anaya et al. [25][5]	144	4	148	
	Xue et al. [11][8]	7	0	7	
	Li et al. [29][1]	15	0	15	
	Chen et al. [26][4]	2	0	2	
	Beltrán-Anaya et al. [24][2]	75	4	79	
	Ajami et al. [23][2]	31	23	54	
	Haddadi et al. [28][6]	35	5	40	
	Farzi et al. [27][7]	38	3	41	286
BC	Chen et al. [26][4]	1	0	1	1
ABD	Xue et al. [11][8]	63	23	86	
	Li et al. [29][1]	28	5	33	
	Chen et al. [26][4]	127	18	145	264
AC	Xue et al. [11][8]	1	0	1	1
ABCCC	Xue et al. [11][8]	1	0	1	
	Farzi et al. [27][7]	1	0	1	2
AB	Beltrán-Anaya et al. [24][3]	22	5	27	
	Ajami et al. [23][2]	25	14	39	66
ABCC	Beltrán-Anaya et al. [25][5]	58	6	64	
	Xue et al. [11][8]	3	0	3	
	Ajami et al. [23][2]	11	21	32	
	Haddadi et al. [28][6]	26	6	32	
	Farzi et al. [27][7]	7	0	7	138
ABBD	Xue et al. [11][8]	24	1	25	
	Chen et al. [26][4]	1	1	1	27
ABBC	Beltrán-Anaya et al. [24][3]	0	1	1	1
Total		749	141	890	

Superscript numbers 1-8 representing the research published order of the clinical research references related to EPIYA motif pattern study and gastric patients.

3 Result And Discussion

3.1 *CagA* EPIYA motif patterns of *H. pylori* strains isolated from infected patients with occurrences of diseases

In Table 2, eight clinical research studies focusing on the EPIYA motif of *CagA* protein provide 10 EPIYA motif patterns, i.e. AD, ABC, BC, ABD, AC, ABCCC, AB, ABCC, ABBD, and ABBC, which *H. pylori* isolated from 890 patients, altogether. EPIYA-ABC patterns were found around 32.13% (286/890) which is the largest proportion. The second, third, and fourth occurrences were EPIYA-ABD, EPIYA-ABCC, and EPIYA-AB patterns which were found around 29.66% (264/890), 15.51% (138/890), and 7.42% (66/890), respectively. Each of EPIYA-BC, EPIYA-AC, and EPIYA-ABBC patterns was rarely found about 0.11% (1/890).

Considering the overall proportion of cancer patients in each EPIYA motif pattern from Table 2, the first rank of EPIYA motif pattern is EPIYA-AB with 28.79% (19/66). The second to fourth ranks are the patterns as follows: EPIYA-ABCC with 23.91% (33/138),

EPIYA-ABD with 17.42% (46/264), EPIYA-ABC with 13.64% (39/286), and EPIYA-ABBD with 7.41% (2/27). On the other hand, EPIYA-BC, EPIYA-ABCCC, and EPIYA-AC patterns are majorly found in gastrointestinal disease patients. But, patterns like EPIYA-AD, EPIYA-BC, EPIYA-AC, EPIYA-ABCCC, and EPIYA-ABBC are still inconclusive, since there are small patient numbers, for now. More clinical experiments for rare pattern cases like theses should be additionally conducted and reported for better insights.

3.2 EPIYA motif patterns of *H. pylori*'s *CagA* protein: NCBI vs clinical data exploration

In order to identify each EPIYA pattern, only the *CagA* protein translated by the first copy of *cagA* gene appearing on its genome was gathered. Denote that the strain 26695 which is a EPIYA-ABC strain was set to a reference as in Xue et al. (2021). Four main motif types, namely, A, B, C, and D were detected in this work, then sequentially read out into EPIYA motif pattern such as ABC patterns. In Table 3, 22 EPIYA motif patterns of 272 *H. pylori* strains were detected. Ten out of 22 patterns were previously reported based on eight previous clinical researches [11, 23-29] in Table 2. However, in this work, twelve patterns were revealed with no previous evidences on the published clinical data based on our present knowledge. They are EPIYA-A, EPIYA-AABC, EPIYA-AABD, EPIYA-AAD, EPIYA-ABABC, EPIYA-ABABD, EPIYA-ABBCC, EPIYA-ABCCCC, EPIYA-ACC, EPIYA-ACCC, EPIYA-ACCCC, and EPIYA-BABD as shown in Table 3.

Table 3. Distribution of 22 EPIYA motif patterns on *CagA* protein of 272 *H. pylori*'s strains derived from NCBI database.

EPIYA Pattern	Number of Occurrences	EPIYA Pattern	Number of Occurrences	EPIYA Pattern	Number of Occurrences
ABD*	115	ABC*	73	ABCC*	19
ACC	11	AD*	9	AB*	6
BC*	5	AC*	5	ABCCC*	4
AABC	4	A	4	AABD	4
ABBD*	2	ABABD	2	AAD	2
ACCC	2	ABBCC	1	ABABC	1
ABCCCC	1	ACCCC	1	ABBC*	1
BABD	1				

* EPIYA patterns also found in *H. pylori*'s clinical isolates from Table 2.

Different types of EPIYA motifs can be found in *H. pylori* isolated from human hosts across various geographic origins [30]. The EPIYA motif can be utilized in studying relationship between variation in EPIYA motifs and disease severity [31]. In our prepared dataset, it is shown that there is no clinical evidence of twelve *CagA* EPIYA patterns based on previous researches, therefore, this is one challenge to fill in missing knowledge.

3.3 Two amino acid positions in *CagA* amino acid polymorphism study

According to Xue et al., 2021, several amino acid polymorphisms in the *H. pylori* sequences around the first EPIYA motif location closing to the *CagA* C-terminal region have been analyzed. Their *H. pylori* sequences were isolated from patients with the known gastric disease. Based on their analysis result, 86 *CagA* proteins with EPIYA-ABD pattern from gastric cancer patients, two-deletion positions before the first EPIYA motif region highly occur significantly. Moreover, deletions including substitution at these two positions are significantly related to gastric cancer, as well. Therefore, in our analysis, two chi-square tests

at 0.05 significant level were computed to observe the two followings. First, the significant difference between two-deletion occurrence numbers of their 86-EPIYA-ABD pattern set and our 115-EPIYA-ABD pattern set. Second, the significant difference between two-deletion including substitution occurrence numbers of these both EPIYA-ABD pattern sets. At 0.05 significant level, there was no significant difference in the two-deletion occurrence numbers ($\chi 2 = 0.103$, p-value = 0.748), in contrast to the deletions including substitution occurrence numbers ($\chi 2 = 11.946$, p-value = 0.0005). In other words, the two-deletion occurrence numbers are quite indifferent unlike the combined occurrence numbers of two-deletion and substitution positions. To explain in detail, our 115-EPIYA-ABD pattern set contains 13 strains (13/115 or 11.3%) with two-deletion position (Fig. 1 and 2(A)), while their clinical data, 86-EPIYA-ABD pattern set, contains 11 isolates (Table 4). This two-deletion occurrence numbers are consistent with each other. But, our 80 strains in 115-EPIYA-ABD pattern set (80/115 or 69.6%) are two-deletion and substitution occurrence numbers (Fig. 2(B)) which are totally different from Xue et al. (2021) clinical dataset containing only 39 isolates. So, in the latter case, it is incomparable. For suitable comparison, more substitution position samples of clinical data should be further collected and investigated. From our two $\chi 2$ test results, one might help to confirm that the observe frequencies of two-deletion positions before the first EPIYA type A segment of the EPIYA-ABD pattern could be potential virulence indicator for estimating the risk of gastric cancer. *CagA* produced by *H. pylori* can be transported into the host's gastric epithelial cells. Then, tyrosine (Y) residues of *CagA* EPIYA motif near C-terminal region can be phosphorylated by host proteins such as Src family kinases (SFKs) like SHP-1 and SHP-2, *CagA* C-terminal Src kinase (*Csk*), phosphatidylinositol 3-kinase (*PI3K*), growth factor receptor- bound proteins 2 and 7 (*Grb2* and *Grb7*), and zonula occludens-1 (*ZO-1*) [32] causing irregular signaling pathways and cell dysfunction which might relate to gastrointestinal disease progress. These two-deletion occurrences may possibly increase the interference degree of host proteins and *CagA*, somehow, yielding a higher disease risk. However, future experiments about this assumption should be further conducted.

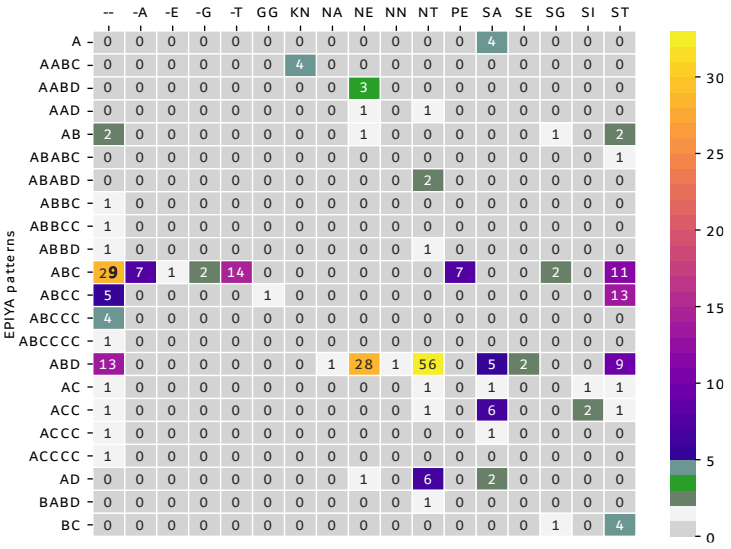

Fig. 1. Distribution of two amino acid positions before the first EPIYA motif segment of each EPIYA pattern on *CagA* protein (272 strains from NCBI database).

Considering 22 EPIYA patterns from our 272-strain set with these two-deletion positions in Fig. 1 and 2, the first and second highest proportions are from EPIYA-ABC (29/73) and

EPIYA-ABD (13/115), respectively. In Fig. 1, two-deletion pattern was detected in *CagA* proteins with EPIYA motif both types C and D. For ST amino acid pattern which is found in a reference sequence of *H. pylori* strain 26695 as well as other patterns starting with S, they were also found in *CagA* proteins with EPIYA motif both types C and D. From our dataset, −A, −E, −G, and −T amino acid patterns which are the first position deletion were only discovered in *CagA* proteins with EPIYA motif type C (i.e. EPIYA-ABC), but −A and −E patterns were found in *CagA* proteins with EPIYA motif type D (i.e. EPIYA-ABD) from another dataset as in Table 4. When focusing on the occurrence number of two-deletion including substitution positions from our strain set (Fig. 2(B)), there is 192 out of 272 strains (70.6%). It is more than three times bigger than the occurrence number of only strains with two-deletion positions (61/272) in Fig. 2(A)).

Table 4. Distribution of two amino acid patterns before the first EPIYA-ABD motif region on *CagA* from our studied dataset with 115 strains and Xue et al. [11] with 86 strains, respectively.

NO	Two-amino acid pattern	Occurrences from our work (115 strains)	Occurrences from Xue et al. [37] (86 strains)
1	--	13	11
2	-A	0	2
3	-E	0	1
4	-G	0	0
5	-T	0	0
6	GG	0	0
7	KN	0	0
8	NA	1	1
9	NE	28	41
10	NN	1	1
11	NT	56	0
12	PE	0	0
13	SA	5	0
14	SE	2	0
15	SG	0	0
16	SI	0	0
17	ST*	9	1
18	N-	0	28

* standing for two amino acids of the reference sequence (*H. pylori* strain 26695) and − standing for deletion.

In Table 4, pattern distribution comparison of these two amino acid positions in *CagA* proteins with EPIYA-ABD pattern from our dataset and Xue et al. (2021) clinical dataset was displayed. Surprisingly, NT, SA, and SE amino acid patterns were only found in our dataset, on the other hand, −A, −E, and N− patterns which are one-deletion were only discovered in Xue et al. (2021) dataset. From Fig. 1, NT and NE amino acid patterns were majorly found in strains with EPIYA motif type D, especially strains with EPIYA-ABD patterns. NE amino acid pattern was the first and second top found patterns in Xue et al. (2021) clinical dataset and our dataset, respectively (as in Table 4). This pattern variation should be further analysis

if there are more data available. For example, NT and NE amino acid pattern distribution in terms of hosts' living subregion in depth, since *H. pylori* strains with EPIYA motif type D were isolated from hosts living in Asia.

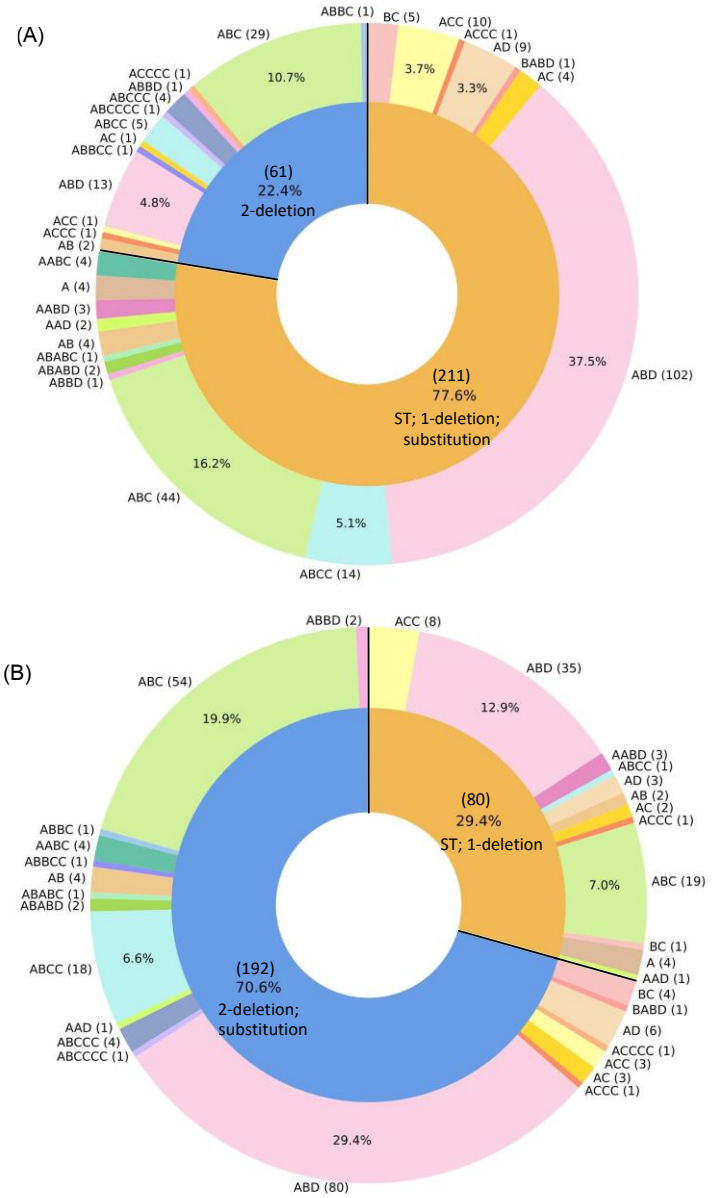

Fig. 2. Distribution of 272 *H. pylori*'s strains derived from NCBI database with their EPIYA motif pattern on *CagA* protein associated with distribution of two amino acid pattern in *CagA* amino acid polymorphism study i.e. two amino acid positions before the first appearance of EPIYA type on EPIYA pattern. (A) Two categories: 1) two-deletion positions and 2) ST pattern, one-deletion, including substitution positions. (B) Two categories: 1) two-deletion including substitution positions. and 2) ST pattern including one-deletion positions.

3.4 Phylogenetic trees of 272 *H. pylori* strains based on *CagA* amino acid and seven-housekeeping gene sequences

Two phylogenetic trees were constructed by different type of data, however, the relationship among strains was revealed by two similar main clades of both trees. These two main clades (pink and blue branches, in Fig. 3 and 4) are two groups of strains containing EPIYA motif types D and C, respectively. There exists 22 EPIYA motif patterns in total (colored circle nodes in Fig. 3 and 4). In Fig. 4, there are 134 and 128 *H. pylori* strains of EPIYA motif types D and C, respectively. The rest ten strains do not belong to either motif type D or C, but they are EPIYA-AB and A motif patterns. Nine out of ten strains were clustered into the EPIYA motif type C clade (the blue branch in Fig. 3 and 4) excepting *H. pylori Hpfe035* strain in Fig. 4.

But, when focusing on both phylogenetic trees in detail, they are slightly different in the term of a position of *H. pylori* strains placing at the end nodes. This variation may be likely attributed to the different data type in tree construction, i.e. nucleotide sequences vs amino acid sequences. Apart from that, the length of each sequence is various. The *CagA* amino acid sequences and the seven-housekeeping gene nucleotide sequences have an average length of 1,181 amino acids and 6,889 bases, respectively.

Generally, a group of strains containing the same *CagA* EPIYA motif type is expected to be clustered into a group of closely related strains based on its genetic background. However, the resulting seven-housekeeping gene tree revealed a small number of strains containing *CagA* EPIYA-AB, ABC, ABCCCC, AC, ACC, and ACCC patterns within EPIYA motif type D clade (the pink branch in Fig. 4). This suggests that the genetic complexity in *H. pylori* strains might relate to variation in the occurrence of their *CagA* EPIYA motif type and pattern.

3.5 Association of *CagA* EPIYA motifs, *cagA* gene copy number and global distribution found in 272 *H. pylori* strains

Focusing on EPIYA motif type D clade of both trees (the pink branch in Fig. 3 and 4), the *cagA* gene of every strain appears only one copy in each genome. The spread of strains with two to four *cagA* copies over the EPIYA motif types C clade displays the quite similar trend on both trees (the blue branch in Fig. 3 and 4). In addition, strains containing four copy numbers of *cagA* gene on their genome, namely, USU101, PMSS1, and 24-A-EK1, are in the same subgroup. Only one strain, MT5125, shows three copy numbers of *cagA* gene on its genome. Besides, the small subtrees on both trees were discovered. They are the light cyan nodes which are strains with *CagA* EPIYA-ABCCC motif pattern indicating that these strains are closely related.

Based on Fig. 3 and 4, considering *H. pylori* strains isolated from human hosts living in many parts of the world, most strains with *CagA* EPIYA motif type D were frequently isolated from human living in Asia. While *H. pylori* strains isolated from human hosts living outside Asia, they mostly contain *CagA* EPIYA motif type C. In general, the EPIYA-ABD motif pattern of 115 strains isolated from human hosts mainly living in Asia is the first highest occurrences (the black nodes on the pink branch in Fig. 3 and 4). These results are similar to previous work [30]. The second highest occurrences are the EPIYA-ABC motif pattern of 73 strains (the dark blue nodes mainly appearing in the blue branch of both trees), which were mostly isolated from human hosts living outside Asia except *H. pylori* Hpfe0021, Santal49, BT302, UBN18 and NP05-124 strains. Denote that strain names with green highlight in the blue branch of Fig. 3 and 4 are the strains isolated from human hosts living in Asia, but most of strains in this clade were isolated from the human hosts living in non-Asia areas. Despite strains with *CagA* EPIYA motif type C being commonly present in hosts from non-Asian

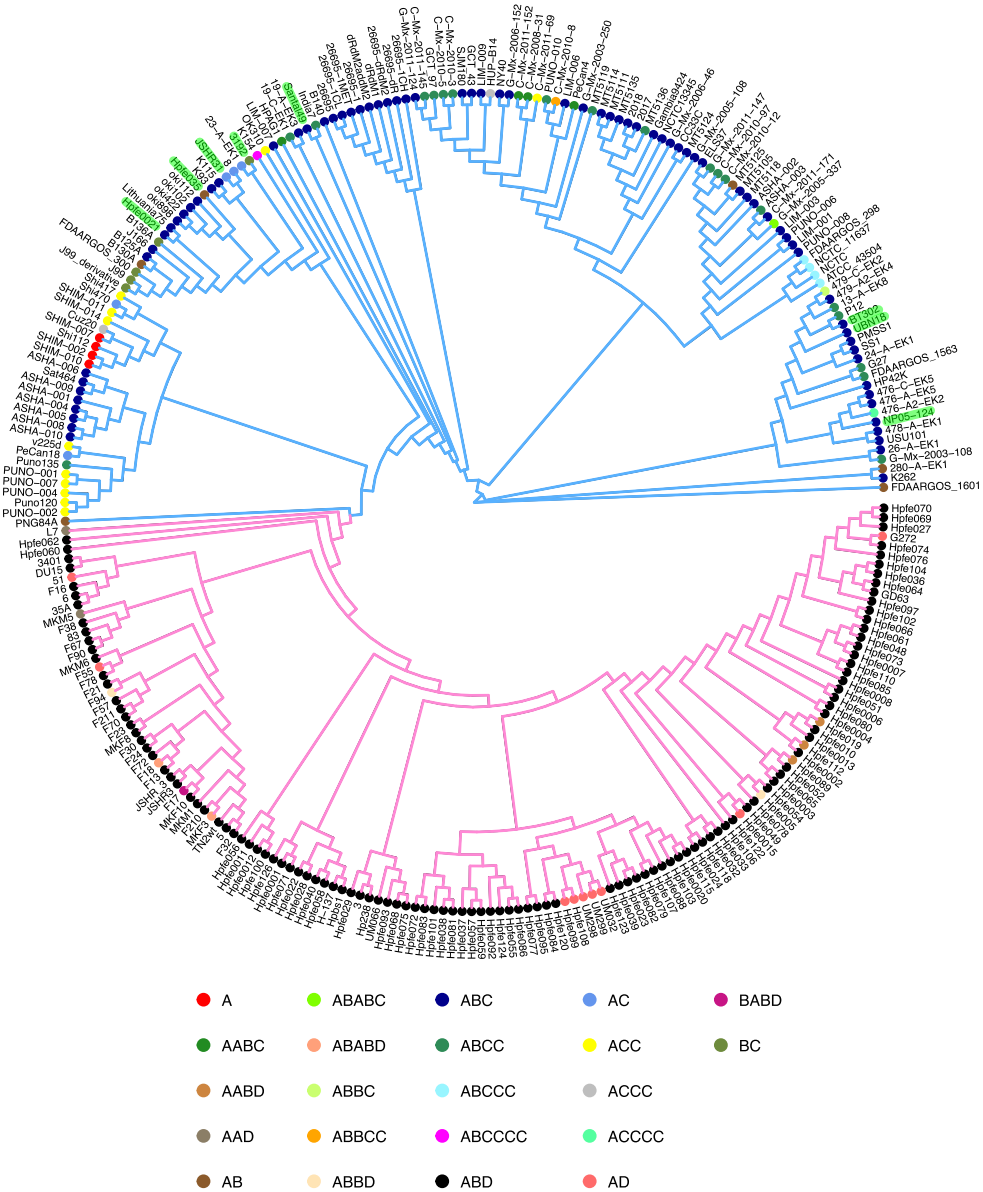

Fig. 3. Phylogenetic tree of *272 H. pylori*'s *CagA* proteins constructed by MEGA7 [18] and displayed by using "ggtree" library version 3.6.2 [19] in R (Section 2.6). Pink and blue branches are two groups of strains containing EPIYA motif types D and C, respectively. Two to four copy numbers of *cagA* gene are labelled in the colored number-in-circle icons.

countries, it can still be found in Asia [33-35]. Such branch occurrences could result from intercontinental migration or colonization in the past, leading to different *H. pylori* strain infections instead of the native strain [35-37].

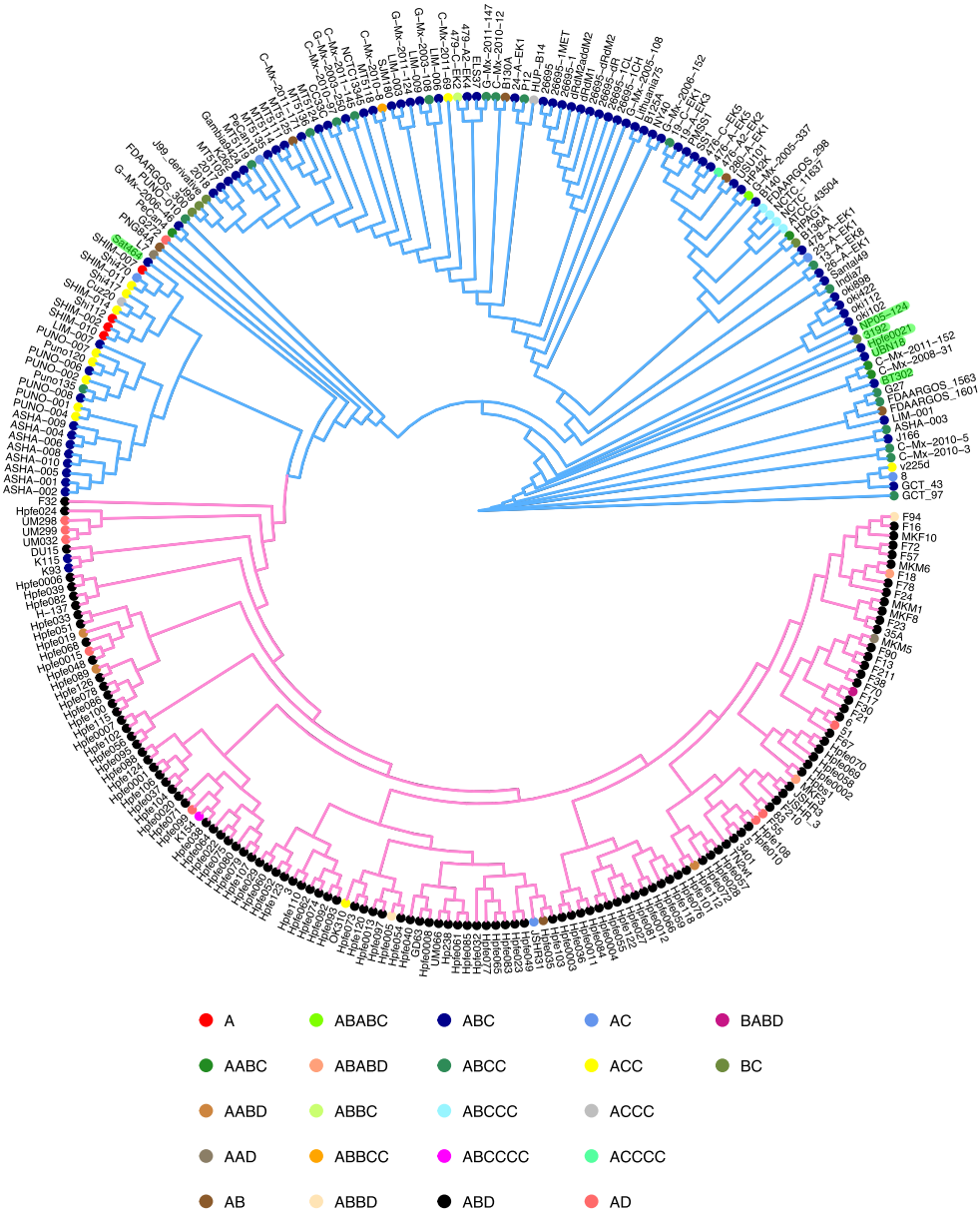

Fig. 4. Phylogenetic tree based on concatenated sequences of seven-housekeeping gene from 272 *H. pylori* strains constructed by MEGA7 [18] and displayed by using "ggtree" library version 3.6.2 [19] in R (Section 2.6). Pink and blue branches are two groups of strains containing EPIYA motif types D and C, respectively. Two to four copy numbers of *cagA* gene are labelled in the colored number-in-circle icons.

3.6 Strains with *CagA* EPIYA-ABD motif pattern linking to both previous patient data and allele sequence data from pubMLST

For each Sequence Type (ST) from the pubMLST database, seven allele sequences were aligned against the downloaded *H. pylori* genome sequences using standalone BLAST. The

best hits with the highest percent identity for the downloaded STs consist of partial sequences of some strain genome sequences from NCBI (as detailed in Section 2.1). A set of 87 STs with its best hits containing partial sequences belonging to only *CagA* EPIYA-ABD strains was focused, since two deletions or substitutions at two amino acid positions before the first EPIYA motif region of EPIYA-ABD strains isolated from gastric patients [11] were previously observed. Additionally, their two-deletion position significantly occurred in the same trend with our studied dataset (as in Section 3.3). Therefore, it is possible to use *H. pylori* partial sequences from PCR products not only to identify *H. pylori*'s ST, but also to estimate severity of the gastric diseases. Due to all of 115 strains with EPIYA-ABD motif pattern containing only one *cagA* gene copy number (Section 3.4, Fig. 3 and 4), in this case, the copy number of the *cagA* gene can not specify a virulence degree for the gastric diseases. This research yielded an initial result based on limited data resources, but it would be more insight and accurate in the future if more studied data available in term of patient's and ST data.

The focused 87 STs where the best hits containing only partial sequences from *H. pylori* EPIYA-ABD strains, generally, not all seven allele sequences were best aligned with sequences from only one same strain. But, there exists 5 STs, i.e. ST numbers 44, 45, 3032, 3034, and 3038 corresponding to only one *H. pylori* strains, i.e. strains F32, DU15, 35A, 83, and F57, respectively, (as in Supplementary Table S2). Their percent identity results were in the range of 96.3 to 100. Although, these 5 STs can be assigned its representative strain, four out of five strains excepting strain F35 contain two amino acids, NT, before the first EPIYA motif which were newly elucidated in this work (as detailed in Table 4 and Supplementary Table S1). So, more future clinical studies may be needed to take strains or *CagA* EPIYA motif patterns into account to gain sufficient gastric disease insights related to *H. pylori*.

The rest 82 out of the 87 STs contain the best hits where the best aligned sequences coming from many strains (Supplementary Table S2). When considering the best hit strains cross all seven allele sequences for these 82 STs, there were 4 out of 82 STs showing that some certain strains were present in all allele sequences, i.e. ST numbers 3035, 3036, 3037, and 3682, corresponding to strains F16, F30, F32, and UM066, respectively. All of their percent identity results were 100. Interestingly, ST numbers 3037 can be identified as the strain F32 with percent identity of 100 found in all seven allele sequences, but the strain F32 was also identified by ST number 44 which all of its best hits came from one strain. Besides, no STs with their best hits were only partial sequences coming from one same strain across five or six allele sequences. Apart from that, the best hits with 100 percent identity of ST number 3036 were partial sequences belonging to the strain F30 across all seven allele sequences. The *H. pylori* strain F30 contains two amino acid deletions before the first EPIYA motif (Supplementary Table S1) which significantly found in patients with gastric cancer [11].

It is noteworthy that few STs can refer to a single strain, but many of them can link to a group of strains with the same EPIYA-ABD motif pattern. Hence, 87 out of 2,740 STs provides the initial information about their possible strains with EPIYA-ABD motif pattern including their pattern of two amino acids before the first EPIYA motif linking to data of previous clinical research. Aside from these 115 EPIYA-ABD strains, computationally, other studied strains with *CagA* can indicate their EPIYA motif pattern, their related STs, and their pattern of two amino acid positions, however more *H. pylori* sequences with *CagA* from gastric patients still be required for meaningful analysis. Previously, the *cagA* gene copy number has been used for evaluating the virulence degree of either *H. pylori* strains or isolates, because several *cagA* gene copies implied to the increasing expression of the toxin protein in the host [38, 39]. But, in our studied dataset, especially a set of *H. pylori* EPIYA-ABD strains, they were isolated from hosts living in Asia and all of them contains only one copy of the *cagA* gene (Supplementary Table S1, Fig. 3 and 4). Hence, another virulence

indicator such as amino acid polymorphism study, e.g. two deletion or substitution positions before the first *CagA* EPIYA motif, which initially studied by Xue et al. (2021) in gastric patient's data was considered, in this work. Moreover, both two amino acid deletion positions and strain identification system like sequence types from the pubMLST database were analyzed together to explore a relation of *H. pylori* EPIYA-ABD strains' sequences and seven allele sequences corresponding to seven housekeeping genes. The results of our study may encourage more future research works about *H. pylori* sequence analysis related to gastric disease severity in terms of important patterns and positions from amino acid polymorphisms, alternative applications of sequence types, and *CagA* EPIYA motif types and patterns.

4 Conclusion

In this study, the 272 *H. pylori* strain's sequence data obtained from the NCBI database were prepared. Each EPIYA motif pattern locating on the *CagA* amino acid sequence and the deletion occurrences at two positions in front of the EPIYA motif type A of the EPIYA-ABD patterns were investigated. The construction of two phylogenetic trees based on the collected seven housekeeping genes and *CagA* amino acid sequence of these 272 strains was also performed. The findings are as follows. First, a study of 272 strain sequences revealed 22 EPIYA motif patterns, while only 10 patterns were found in a compilation of the previous eight clinical studies. Second, in the *CagA* proteins containing EPIYA-ABD pattern, our outcome about the distribution of two deletion positions before the first EPIYA-ABD motif agreed with the previous clinical research dataset. The presence of these two-deletion positions might serve as an indicator of a connection with gastrointestinal diseases and gastric cancer. Especially, in a case of strains with one copy number of *cagA* gene such as strains with *CagA* EPIYA-ABD motif pattern which mostly found in hosts living in Asia, two-deletion amino acid polymorphism study might help to imply virulence level of gastric diseases. Third, two resulting similar phylogenetic trees, con- structed by the 272 *H. pylori* seven-housekeeping gene sequences and the *CagA* proteins showed two main strain groups, i.e. strains with EPIYA motif types C and D, respectively. Finally, our studied *H. pylori* data of 115 strains with *CagA* EPIYA-ABD motif pattern can refer to 87 sequence types (STs) of the pubMLST database via the best BLAST results. Therefore, through allele sequence analysis where data derived from PCR results, species identification system like pubMLST sequence types could be beneficial for detecting the groups of *H. pylori* strains with the same *CagA* EPIYA motif. In summary, combining amino acid polymorphism study and *H. pylori* sequence type identification, which were linked to its *CagA* EPIYA motif pattern, could be used to evaluate the severity of gastric diseases or cancer in the patients infected by *H. pylori* as well as to further develop an effective treatment plan. Two ideas for future research, one could explore the impact of different EPIYA motifs on disease virulence and progression which potentially leads to investigating variously specific therapies. Another research is assessing the effect of genetic variation on treatment outcomes which may give more understanding in the term of personalized medicine for gastric diseases caused by *H. pylori*.

References

1. IARC, IARC Monographs on the Evaluation of Carcinogenic Risks to Humans. *Schistosomes, Liver Flukes and Helicobacter pylori*. Vol. 61. IARC, Lyon, France (1994).
2. I.J. Choi, C.G. Kim, J.Y. Lee, Y.I. Kim, M.C. Kook, B. Park, J. Joo, Family history of gastric cancer and *Helicobacter pylori* treatment. *N. Engl. J. Med.* **382**, 427–436 (2020). https://doi.org/10.1056/NEJMoa1909666

3. J.C. Caguazango, Ecological models of gastric microbiota dysbiosis: *Helicobacter pylori* and gastric carcinogenesis. *Medicine in Microecology* **3**, 100010 (2020). https://doi.org/10.1016/j.medmic.2020.100010
4. M. Achtman, T. Azuma, Y. Berg, G. Morelli, Z.J. Pan, S. Suerbaum, S.A. Thompson, A. Van Der Ende, L.J. Van, Recombination and clonal groupings within *Helicobacter pylori* from different geographical regions. *Mol. Microbiol.* **32**, 459–470 (1999). https://doi.org/10.1046/j.1365-2958.1999.01382.x
5. K.A. Jolley, J.E. Bray, M.C.J. Maiden, Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website, and their applications. *Wellcome Open Res.* **3**, 124 (2018). https://doi.org/10.12688/wellcomeopenres.14826.1
6. S. Mendoza-Elizalde, A.C. Cortés-Márquez, G. Zuñiga, R. Cerritos, P. Valencia-Mayoral, A.C. Sánchez, H. Olivares-Clavijo, N. Velázquez-Guadarrama, Inference from the analysis of genetic structure of *Helicobacter pylori* strains isolates from two pediatric patients with recurrent infection. *BMC Microbiol.* **19**, 184 (2019). https://doi.org/10.1186/s12866-019-1554-z
7. N. Raaf, W. Amhis, H. Saoula, A. Abid, M. Nakmouche, A. Balamane, N.A. Arous, M. Ouar-Korichi, F.F. Vale, L. Bénéjat, F. Mégraud, Prevalence, antibiotic resistance, and MLST typing of *Helicobacter pylori* in Algiers, Algeria. *Helicobacter* **22**, e12446 (2017). https://doi.org/10.1111/hel.12446
8. N.J. Foegeding, R.R. Caston, M.S. McClain, M.D. Ohi, T.L. Cover, An overview of *Helicobacter pylori* VacA toxin biology. *Toxins (Basel)* **8**, 173 (2016). https://doi.org/10.3390/toxins8060173
9. X.-Y. Yuan, J.-J. Yan, Y.-C. Yang, C.-M. Wu, Y. Hu, J.-L. Geng, *Helicobacter pylori* with East Asian-type cagPAI genes is more virulent than strains with Western-type in some cagPAI genes. *Braz. J. Microbiol.* **48**, 218–224 (2017). https://doi.org/10.1016/j.bjm.2016.12.00
10. S. Diechler, B.E. Chichirau, G. Posselt, D.N. Sgouras, S. Wessler, *Helicobacter pylori CagA* EPIYA motif variations affect metabolic activity in B cells. *Toxins (Basel)* **13**, 592 (2021). https://doi.org/10.3390/toxins13090592
11. Z. Xue, Y. You, L. He, Y. Gong, L. Sun, X. Han, R. Fan, K. Zhai, Y. Yang, M. Zhang, X. Yan, J. Zhang, Diversity of 3' variable region of *cagA* gene in *Helicobacter pylori* strains isolated from Chinese population. *Gut Pathog.* **13**, 23 (2021). https://doi.org/10.1186/s13099-021-00419-3
12. K. Thorell, Z.Y. Muñoz-Ramírez, D. Wang, S. Sandoval-Motta, R.B. Agostini, S. Ghirotto, R.C. Torres, HpGP Research Network, D. Falush, M.C. Camargo, C.S. Rabkin, The *Helicobacter pylori* Genome Project: insights into *H. pylori* population structure from analysis of a worldwide collection of complete genomes. *Nat. Commun.* **14**, 8184 (2023). https://doi.org/10.1038/s41467-023-43562-y
13. M. Khaledi, N. Bagheri, M. Validi, B. Zamanzad, H. Afkhami, J. Fathi, G. Rahimian, A. Gholipour, Determination of *CagA* EPIYA motif in *Helicobacter pylori* strains isolated from patients with digestive disorder. *Heliyon* **6**, e04971 (2020). https://doi.org/10.1016/j.heliyon.2020.e04971
14. K.S. Papadakos, I.S. Sougleri, A.F. Mentis, E. Hatziloukas, D.N. Sgouras, Presence of terminal EPIYA phosphorylation motifs in *Helicobacter pylori CagA* contributes to IL-8 secretion, irrespective of the number of repeats. *PLoS One* **8**, e56291 (2013). https://doi.org/10.1371/journal.pone.0056291
15. M. Keikha, M. Karbalaei, EPIYA motifs of *Helicobacter pylori cagA* genotypes and gastrointestinal diseases in the Iranian population: a systematic review and meta-

analysis. *New Microbes New Infect.* **41**, 100865 (2021). https://doi.org/10.1016/j.nmni.2021.100865

16. A.C.E. Darling, B. Mau, F.R. Blattner, N.T. Perna, Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* **14**, 1394–1403 (2004). https://doi.org/10.1101/gr.2289704
17. The MathWorks Inc., MATLAB. The MathWorks Inc, Natick, Massachusetts, United States of America. Version R2022a (2022).
18. S. Kumar, G. Stecher, K. Tamura, MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* **33**, 1870–1874 (2016). https://doi.org/10.1093/molbev/msw054
19. G. Yu, D.K. Smith, H. Zhu, Y. Guan, T.T.-Y. Lam, ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017). https://doi.org/10.1111/2041-210X.12628
20. J.D. Thompson, D.G. Higgins, T.J. Gibson, CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680 (1994). https://doi.org/10.1093/nar/22.22.4673
21. K. Tamura, M. Nei, S. Kumar, Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 11030–11035 (2004). https://doi.org/10.1073/pnas.0404206101
22. N. Saitou, M. Nei, The Neighbor-Joining Method: A New Method for Reconstructing Phylogenetic Trees. *Mol. Biol. Evol.* **4**, 406–425 (1987). https://doi.org/10.1093/oxfordjournals.molbev.a040454
23. A. Ajami, M. Shadman, A. Rafiei, V. Hosseini, A. TalebiBezmin Abadi, A. Alizadeh, Z. Hosseini-khah, Prevalence of EPIYA motifs in *Helicobacter pylori* strains isolated from patients with gastroduodenal disorders in northern Iran. *Res. Mol. Med.* **1**, 29–34 (2013). https://doi.org/10.18869/acadpub.rmm.1.1.29
24. F.O. Beltrán-Anaya, Diversidad de la región variable 3 del oncogén *cagA* de *H. pylori* en pacientes con gastritis crónica y cáncer gástrico. Master's thesis, Universidad Autónoma de Guerrero (México) (2013).
25. F.O. Beltrán-Anaya, T.M. Poblete, A. Román-Román, S. Reyes, J. de Sampedro, O. Peralta-Zaragoza, M.Á. Rodríguez, O. del Moral-Hernández, B. Illades-Aguiar, G. Fernández-Tilapa, The EPIYA-ABCC motif pattern in *CagA* of *Helicobacter pylori* is associated with peptic ulcer and gastric cancer in the Mexican population. *BMC Gastroenterol.* **14**, 223 (2014). https://doi.org/10.1186/s12876-014-0223-9
26. C.Y. Chen, F.Y. Wang, H.J. Wan, X.X. Jin, J. Wei, Z.K. Wang, C. Liu, H. Lu, H. Shi, D.H. Li, J. Liu, Amino acid polymorphisms flanking the EPIYA-A motif of *Helicobacter pylori CagA* C-terminal region are associated with gastric cancer in East China: experience from a single center. *J. Dig. Dis.* **14**, 358–365 (2013). https://doi.org/10.1111/1751-2980.12056
27. N. Farzi, A. Yadegar, H.A. Aghdaei, Y. Yamaoka, M.R. Zali, Genetic diversity and functional analysis of *oipA* gene in association with other virulence factors among *Helicobacter pylori* isolates from Iranian patients with different gastric diseases. *Infect. Genet. Evol.* **60**, 26–34 (2018). https://doi.org/10.1016/j.meegid.2018.02.017
28. M.H. Haddadi, A. Bazargani, R. Khashei, M.R. Fattahi, K.B. Lankarani, M. Moini, S.M.H.R. Hosseini, Different distribution of *Helicobacter pylori* EPIYA-*cagA* motifs and *dupA* genes in the upper gastrointestinal diseases and correlation with clinical

outcomes in Iranian patients. *Gastroenterol. Hepatol. Bed Bench* **8** (Suppl 1), S37–46 (2015).

29. J. Li, Z. Ou, F. Wang, Y. Guo, R. Zhang, J. Zhang, P. Li, W. Xu, Y. He, Distinctiveness of the *cagA* genotype in children and adults with peptic symptoms in South China. *Helicobacter* **14**, 248–255 (2009). https://doi.org/10.1111/j.1523-5378.2009.00690.x
30. J. Lind, S. Backert, K. Pfleiderer, D.E. Berg, Y. Yamaoka, H. Sticht, N. Tegtmeyer, Systematic analysis of phosphotyrosine antibodies recognizing single phosphorylated EPIYA motifs in *CagA* of Western-type *Helicobacter pylori* strains. *PLoS One* **9**, e96488 (2014). https://doi.org/10.1371/journal.pone.0096488
31. H. Higashi, R. Tsutsumi, A. Fujita, S. Yamazaki, M. Asaka, T. Azuma, M. Hatakeyama, Biological activity of the *Helicobacter pylori* virulence factor *CagA* is determined by variation in the tyrosine phosphorylation sites. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 14428–14433 (2002). https://doi.org/10.1073/pnas.222375399
32. M. Selbach, F.E. Paul, S. Brandt, P. Guye, O. Daumke, S. Backert, C. Dehio, M. Mann, Host cell interactome of tyrosine-phosphorylated bacterial proteins. *Cell Host Microbe* **5**, 397–403 (2009). https://doi.org/10.1016/j.chom.2009.03.004
33. S.S. Duncan, P.L. Valk, M.S. McClain, C.L. Shaffer, J.A. Metcalf, S.R. Bordenstein, T.L. Cover, Comparative genomic analysis of East Asian and non-Asian *Helicobacter pylori* strains identifies rapidly evolving genes. *PLoS One* **8**, e55120 (2013). https://doi.org/10.1371/journal.pone.0055120
34. J. Lind, S. Backert, R. Hoffmann, J. Eichler, Y. Yamaoka, G.I. Perez-Perez, J. Torres, H. Sticht, N. Tegtmeyer, Systematic analysis of phosphotyrosine antibodies recognizing single phosphorylated EPIYA motifs in *CagA* of East Asian-type *Helicobacter pylori* strains. *BMC Microbiol.* **16**, 201 (2016). https://doi.org/10.1186/s12866-016-0820-6
35. S. Sahara, M. Sugimoto, R.-K. Vilaichone, V. Mahachai, H. Miyajima, T. Furuta, Y. Yamaoka, Role of *Helicobacter pylori cagA* EPIYA motif and *vacA* genotypes for the development of gastrointestinal diseases in Southeast Asian countries: a meta-analysis. *BMC Infect. Dis.* **12**, 223 (2012). https://doi.org/10.1186/1471-2334-12-223
36. B.A. Salih, B.K. Bolek, M.T. Yildiz, S. Arikan, Phylogenetic analysis of *Helicobacter pylori cagA* gene of Turkish isolates and the association with gastric pathology. *Gut Pathog.* **5**, 33 (2013). https://doi.org/10.1186/1757-4749-5-33
37. K. Tissera, M.-A. Kim, J. Lai, S. Angulmaduwa, A. Kim, D.S. Merrell, J.-H. Kim, H. Su, J.-H. Cha, Characterization of East-Asian *Helicobacter pylori* encoding Western EPIYA-ABC *CagA*. *J. Microbiol.* **60**, 207–214 (2022). https://doi.org/10.1007/s12275-022-1483-7
38. S. Jang, L.M. Hansen, H. Su, J.V. Solnick, J.-H. Cha, Host immune response mediates changes in *cagA* copy number and virulence potential of *Helicobacter pylori*. *Gut Microbes* **14**, 2044721 (2022). https://doi.org/10.1080/19490976.2022.2044721
39. P. Saniee, S. Jalili, P. Ghadersoltani, L. Daliri, F. Siavoshi, Individual hosts carry *H. pylori* isolates with different *cagA* features - motifs and copy number. *Infect. Genet. Evol.* **93**, 104961 (2021). https://doi.org/10.1016/j.meegid.2021.104961

Appendix

Supplementary Materials: https://github.com/Narkwichearn/HpyloriABD stored Tables S1 and S2.